A Probabilistic Annotation Model for Crowdsourcing Coreference

Silviu Paun[1], Jon Chamberlain[2], Udo Kruschwitz[2], Juntao Yu[1] and Massimo Poesio[1]
[1]School of Electronic Engineering and Computer Science, Queen Mary University of London
[2]School of Computer Science and Electronic Engineering, University of Essex

Abstract

The availability of large scale annotated corpora for coreference is essential to the development of the field. However, creating resources at the required scale via expert annotation would be too expensive. Crowdsourcing has been proposed as an alternative; but this approach has not been widely used for coreference. This paper addresses one crucial hurdle on the way to make this possible, by introducing a new model of annotation for aggregating crowdsourced anaphoric annotations. The model is evaluated along three dimensions: the accuracy of the inferred mention pairs, the quality of the post-hoc constructed silver chains, and the viability of using the silver chains as an alternative to the expert-annotated chains in training a state of the art coreference system. The results suggest that our model can extract from crowdsourced annotations coreference chains of comparable quality to those obtained with expert annotation.

1 Introduction

The task of identifying and resolving anaphoric reference to discourse entities, known in NLP as coreference resolution, has long been considered a core aspect of language interpretation (Poesio et al., 2016b), also because of its role in applications such as summarization (Baldwin and Morton, 1998; Steinberger et al., 2007), information extraction (Humphreys et al.) or question answering (Morton, 1999; Zheng, 2002).

In the 1990s the field made a paradigmatic turn towards corpus based approaches initiated by campaigns such as MUC (Grishman and Sundheim, 1995; Chinchor, 1998) and since then we have seen the development of a range of data-driven approaches, spurred by the development of ever larger and richer datasets. Nowadays, a variety of datasets exist for several languages (Poesio et al., 2016a). These include medium-scale multilingual datasets such as ONTONOTES (Pradhan et al., 2007; Weischedel et al., 2011), which led to the most recent evaluation campaigns, in particular CONLL 2012 (Pradhan et al., 2012), and are used in most current research (Björkelund and Kuhn, 2014; Martschat and Strube, 2015; Clark and Manning, 2016; Lee et al., 2017). However, there are still many languages and domains for which no such resources are available, and even for English much larger corpora than ONTONOTES will eventually be required.

However, annotating data on the scale required to train state of the art systems using traditional expert annotation would be unaffordable. One alternative is to employ **crowdsourcing**, either via platforms like Amazon Mechanical Turk and Crowdflower, or using **Games-With-A-Purpose** (Poesio et al., 2017). Studies such as (Snow et al., 2008; Raykar et al., 2010) have shown that when a sufficiently large number of workers is employed, expert-level quality can be achieved, at a fraction of the cost required to create such resources using traditional methods. The one effort to create a large-scale coreference corpus entirely through crowdsourcing, the *Phrase Detectives* project (Poesio et al., 2013; Chamberlain et al., 2016; Chamberlain, 2016), employs the *Phrase Detectives* game with a purpose. The *Phrase Detectives* corpus consists of 843 documents for a total of 1.2 million tokens and 392,741 markables; at present, 563 documents for a total of 360,000 tokens have been annotated.[1] A second coreference corpus created using crowdsourcing (in the context of a trivia game) also exists, the

[1]Note that although the *Phrase Detectives* corpus is slightly smaller in terms of tokens than the currently largest coreference corpus for English, the CONLL 2012 dataset (Pradhan et al., 2012), it has about twice the number of markables, 390,000 vs. 190,000.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1926–1937
Brussels, Belgium, October 31 - November 4, 2018.

Quiz Bowl dataset (Guha et al., 2015).[2]

However, such existing corpora are not widely used yet. One of the reasons for this is the lack of suitable **aggregation methods** for anaphora. Crowdsourced annotations require aggregation methods to select among the different interpretations produced by the crowd. Standard practice for crowdsourced data analysis has seen a shift in recent years from simple majority vote to much more effective aggregation methods (Smyth et al., 1994; Quoc Viet Hung et al., 2013; Sheshadri and Lease, 2013; Carpenter, 2008; Hovy et al., 2013; Passonneau and Carpenter, 2014). Probabilistic models of annotation, in particular, make it possible to characterize the accuracy of the annotators and correct for their bias (Dawid and Skene, 1979; Passonneau and Carpenter, 2014), to account for item-level effects (e.g.: difficulty) (Whitehill et al., 2009), and to employ different pooling strategies (Carpenter, 2008). However, existing models of annotation cannot be used for anaphora. Such methods assume that coders choose between a fixed set of general labels, the same labels across all annotated items. In anaphoric annotation, by contrast, coders relate markables to coreference chains which depend on the markables that are annotated in that given document (Passonneau, 2004; Artstein and Poesio, 2008)

Contributions In this paper we propose a mention pair-based approach to aggregating crowdsourced anaphoric annotations. Concretely, we introduce a new model of annotation capable of inferring the most likely mention pairs from crowd-annotated anaphoric relations. We then use these pairs to build the most likely coreference chains. This approach to building chains is evaluated on both crowdsourced and synthetic (via simulation) coreference datasets. The evaluations include assessing the accuracy of the inferred mention pairs; the quality of the chains; and the viability of using these chains derived from mention pairs as an alternative to gold chains when training a state of the art coreference system. We conclude by also demonstrating the quality of the proposed model in a standard annotation task. The implementation is available as supplementary material.

[2]Another corpus creation project using crowdsourcing (and also games) for anaphoric annotation is the *Groningen Meaning Bank* (Bos et al., 2017). However, in the GMB crowdsourcing is not used to generate interpretations: players correct automatically annotated interpretations rather than providing the annotations themselves. Another crucial difference is that interpretations are not aggregated in the sense discussed below; rather, an expert adjudicates between the interpretations produced by players.

2 A Mention-Pair Model of Annotation

Traditional models of annotation (Dawid and Skene, 1979; Smyth et al., 1994; Raykar et al., 2010; Hovy et al., 2013) are specified assuming the annotations are chosen among a general set of classes that is consistent across the annotated items. This is the case in a type of annotation closely related to anaphoric annotation, **information status** annotation (Nissim et al., 2004; Riester et al., 2010). In this type of annotation, an annotator marks a mention as either **discourse old** (DO) – referring to an existing entity (coreference chain) – or as **discourse-new** (DN) – introducing a new coreference chain, but without specifying *which* coreference chain the mention belongs to, if any. We will refer below to categories such as DN and DO as **(general) classes**.

Traditional models of annotation can model this type of annotation, but not the task of anaphoric annotation proper. In standard annotation schemes for anaphora/coreference (Poesio et al., 2016a) the annotator may mark a mention as referring to a discourse new entity as above; but in case the mention is identified as discourse-old, this entity, or **coreference chain**–the set of coreferring mentions–is also specified. The available coreference chains differ from document to document.

Our proposal for a probabilistic model of this type of annotation is based on one of the most widely used models of coreference *resolution*: the **mention pair** model. In the mention pair model, the task of linking the mention to a coreference chain/entity is split in two parts: classifying mention pairs as coreferring or not, and subsequent clustering (Soon et al., 2001; Hoste, 2016). The model we propose addresses the first part.

More formally, the crowdsourced data to be modeled consists of I mentions (indexed by i) annotated by a total of J coders (indexed by j). Each mention i has N_i annotations (indexed by n), for a total of M_i distinct **labels** (indexed by m). Each label m of mention i belongs to a **class** $z_{i,m}$. The label of a mention could be the ID of the antecedent, in case that mention is annotated as belonging to the discourse old (general) class; or could be discourse new or another general class (e.g.: property, non referring). In these latter cases, the labels coincide with the classes they

belong to.

An important difficulty we had to address is label sparsity. The solution we propose is to transform the mention-level annotations into a series of binary decisions with respect to each candidate label. In the extended literature this is often referred to as the binary relevance method (Tsoumakas and Katakis, 2007; Madjarov et al., 2012). We then model these (label-level) decisions as the result of the sensitivity (the true positive rate) and specificity (the true negative rate) of the annotators which we assume are class dependent. This latter assumption allows inferring different levels of annotator ability for each class (e.g.: capturing that DO labels are generally harder compared to DN).

The graphical model of our **Mention Pair Annotations** model (MPA) is presented in Figure 1, while the generative process is given below:

- For every class $h \in \{1, 2, ..., K\}$:
  - Draw class specific true label likelihood $\pi_h \sim \mathsf{Beta}(a, b)$
- For every annotator $j \in \{1, 2, ..., J\}$:
  - For every class $h \in \{1, 2, ..., K\}$:
    - Draw sensitivity $\alpha_{j,h} \sim \mathsf{Beta}(d, e)$
    - Draw specificity $\beta_{j,h} \sim \mathsf{Beta}(t, u)$
- For every mention $i \in \{1, 2, ..., I\}$:
  - For every candidate label $m \in \{1, 2, ..., M_i\}$:
    - Draw true label indicator $c_{i,m} \sim \mathsf{Bern}(\pi_{z_{i,m}})$
    - For every position $n \in \{1, 2, ..., N_i\}$:
      - If $c_{i,m} = 1$ then draw decision $y_{i,m,n} \sim \mathsf{Bern}(\alpha_{jj[i,m,n],z_{i,m}})$[3]
      - Otherwise, draw decision $y_{i,m,n} \sim \mathsf{Bern}(1 - \beta_{jj[i,m,n],z_{i,m}})$

The model addresses the first part of the mention pair framework: the posterior of the true label indicators is used to link each mention with the most likely label, obtaining the mention pairs. The coreference chains are then built by following the link structure from the inferred pairs.

Note that for a traditional annotation task with no distinction between generic classes and specific labels the MPA model is equivalent to training K binary Bayesian versions of the Dawid and Skene (1979) model (one for each general class) on data processed using the binary relevance method. Note also that whereas traditional models of annotation assume one true class per annotated item, an implicit benefit of our approach is allowing for potentially multiple true classes, which can be useful to detect ambiguity (Poesio and Artstein, 2005), but we don't exploit that in this work.

[3]Notation: jj[i,m,n] returns the index of the annotator who made the n-th decision on the m-th label of mention i.

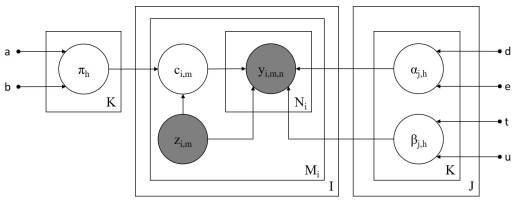

Figure 1: Plate diagram for MPA

2.1 Parameter Estimation

We infer the parameters of the proposed model using Variational Inference (VI). Unlike Markov Chain Monte Carlo (MCMC) approaches (e.g.: Gibbs Sampling, Hamiltonian Monte Carlo), VI is deterministic, fast, and benefits from a clear convergence criterion (Blei et al., 2017).

Specifically we approximate the intractable posterior $p(\theta|D)$ with a variational distribution $q(\theta)$ such that the Kullback-Leibler (KL) divergence between the two distributions is minimized. It can be shown this minimization is equivalent to maximizing the evidence lower bound (ELBO) below:

$$\begin{aligned}\mathcal{L} = E_q[\log p(\pi, \alpha, \beta, c, y|a, b, d, e, t, u, z)] \\ - E_q[\log q(\pi, \alpha, \beta, c|\lambda, \eta, \gamma, \mu, \theta, \epsilon, \phi, \zeta)]\end{aligned} \quad (1)$$

We need a variational distribution q that is tractable under expectations. Following common practice (Blei et al., 2003; Hoffman et al., 2013; Blei et al., 2017), we choose q to be in the mean field variational family where each hidden variable is independent and governed by its own parameter. Elegant solutions have been derived for models whose complete conditionals are in the exponential family (Blei and Jordan, 2006; Hoffman et al., 2013). Concretely, we used the fact that the natural parameters of the variational distributions are equal to the expected value of the natural parameters of the corresponding complete conditionals.

The derivations are standard in the VI literature (see, for example, Hoffman et al., 2013). (To save space, we only provide here the update formulas of

the variational parameters; supplementary details are in the Appendix.)

Equations (2) and (3) give the variational update formulas for the class-level true label likelihood. We have $q(\pi_h|\lambda_h, \eta_h) = \mathsf{Beta}(\lambda_h, \eta_h)$, where:

$$\lambda_h = a + \sum_{i,m}^{I,M_i} I(z_{i,m} = h) E_q[I(c_{i,m} = 1)] \quad (2)$$

$$\eta_h = b + \sum_{i,m}^{I,M_i} I(z_{i,m} = h) E_q[I(c_{i,m} = 0)] \quad (3)$$

In Equation (4) and (5) we list the variational update formulas for the class-level annotator sensitivity. We have $q(\alpha_{j,h}|\gamma_{j,h}, \mu_{j,h}) = \mathsf{Beta}(\gamma_{j,h}, \mu_{j,h})$, where:

$$\gamma_{j,h} = d + \sum_{i,m,n}^{I,M_i,N_i} I(jj[i,m,n] = j) \; I(z_{i,m} = h) I(y_{i,m,n} = 1) E_q[I(c_{i,m} = 1)] \quad (4)$$

$$\mu_{j,h} = e + \sum_{i,m,n}^{I,M_i,N_i} I(jj[i,m,n] = j) \; I(z_{i,m} = h) I(y_{i,m,n} = 0) E_q[I(c_{i,m} = 1)] \quad (5)$$

In Equations (6) and (7) we list the variational update formulas for the class-level annotator specificity. We have $q(\beta_{j,h}|\theta_{j,h}, \epsilon_{j,h}) = \mathsf{Beta}(\theta_{j,h}, \epsilon_{j,h})$, where:

$$\theta_{j,h} = t + \sum_{i,m,n}^{I,M_i,N_i} I(jj[i,m,n] = j) \; I(z_{i,m} = h) I(y_{i,m,n} = 0) E_q[I(c_{i,m} = 0)] \quad (6)$$

$$\epsilon_{j,h} = u + \sum_{i,m,n}^{I,M_i,N_i} I(jj[i,m,n] = j) \; I(z_{i,m} = h) I(y_{i,m,n} = 1) E_q[I(c_{i,m} = 0)] \quad (7)$$

In Equations (8) and (9) we list the variational update formulas for the true label indicator. We have $q(c_{i,m}|\phi_{i,m}) = \mathsf{Bern}(\phi_{i,m})$, where $\zeta_{i,m} = 1 - \phi_{i,m}$ and:

$$\log \phi_{i,m} \propto E_q[\log \pi_{z_{i,m}}] + \sum_{n=1}^{N_i} I(y_{i,m,n} = 1) E_q[\log \alpha_{jj[i,m,n],z_{i,m}}] + I(y_{i,m,n} = 0) E_q[\log(1 - \alpha_{jj[i,m,n],z_{i,m}})] \quad (8)$$

$$\log \zeta_{i,m} \propto E_q[\log(1 - \pi_{z_{i,m}})] + \sum_{n=1}^{N_i} I(y_{i,m,n} = 0) E_q[\log \beta_{jj[i,m,n],z_{i,m}}] + I(y_{i,m,n} = 1) E_q[\log(1 - \beta_{jj[i,m,n],z_{i,m}})] \quad (9)$$

Finally, for the above formulas, we used the fact that $E_q[I(c_{i,m} = 1)] = \phi_{i,m}$. The other expectations can be easily calculated noting that for a distribution part of the exponential family, the first derivative of the log normalizer is equal to the expected value of the sufficient statistics (Blei et al., 2003). For example, $E_q[\log \pi_{z_{i,m}}] = \Psi(\lambda_{z_{i,m}}) - \Psi(\lambda_{z_{i,m}} + \eta_{z_{i,m}})$, where $\Psi(.)$ is the digamma function. Similar observations apply to the α and β related expectations.

The algorithm, known as Coordinate Ascent Variational Inference (CAVI) (Blei et al., 2017), involves iterating between Equations (2), (3), (4), (5), (6), (7), (8) and (9) until convergence. The ELBO expressed in Equation (1) is guaranteed to increase at every step. Convergence is achieved when the ELBO plateaus. Throughout the experiments we used non-informative, uniform priors.

3 Evaluation

We carried out a series of evaluations of increasing complexity of our MPA model. We first assess the accuracy of the inferred mention pairs. Second, we cluster the pairs into appropriate coreference chains and evaluate the quality of these chains. Third, we assess the viability of using silver chains as an alternative to the gold chains when training a state of the art coreference system. Finally, we conclude the evaluation with a performance check in a standard annotation task.

3.1 Datasets

The largest coreference dataset with crowdsourced annotations is the *Phrase Detectives* corpus. A subset of this corpus is the *Phrase Detectives 1.0* dataset (Chamberlain et al., 2016), which also includes gold annotations and can therefore be used to evaluate the accuracy of MPA at mention-pair and coreference chain inference, but is too small to train a state-of-the-art coreference system. To carry out this second type of evaluation we used the approach, common in the crowdsourcing literature (Carpenter, 2008; Raykar et al., 2010; Hovy et al., 2013; Felt et al., 2014), of generating simulated datasets by corrupting the gold standard

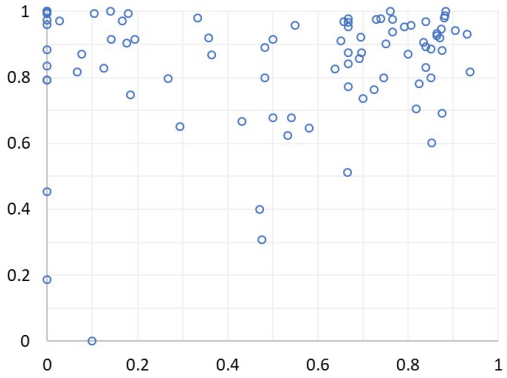

Figure 2: Sensitivity profiles extracted from the PD corpus: DO (x-axis) vs. DN (y-axis)

of an existing corpus. For this purpose, we use the CONLL-2012 dataset (Pradhan et al., 2012), at present the standard dataset for coreference resolution.

3.1.1 Crowdsourced Data

The *Phrase Detectives (PD) 1.0* dataset has been annotated using the *Phrase Detectives* game with a purpose.[4] The annotation scheme for PD is based on that for the ARRAU corpus (Poesio et al., 2018). Players have to label predefined[5] markables with one of the following categories: **non-referring** (e.g., for expletives), **discourse-new**, **discourse-old** (in which case an antecedent is also marked, the most recent mention belonging to the antecedent's coreference chain), or **property** (for appositions and copular structures). The PD 1.0 dataset is the portion of the corpus that contains, in addition to the annotations by the players, a gold label for each markable. The coreference chains are obtained using a simple clustering of the mention pairs. An important limitation of this corpus is its small size (around 6000 markables from 45 documents), making it unfit for the training and evaluation of state of the art supervised systems.

3.1.2 Synthetic Data

The CONLL-2012 dataset specifies gold chains, not mention pairs. So we need first to extract appropriate mention pairs from these chains. To do this, for each mention we select as gold label the closest mention from its gold chain (or discourse new if the mention is the first in its chain).

[4] http://www.phrasedetectives.org

[5] In standard annotation projects markables are predefined for better agreement. The markables used in PD are automatically identified, but players can highlight errors in markable identification that can then be corrected.

Simulation	Profile Type	Error Distribution
1	Synthetic	Uniform
2	Synthetic	Sparse
3	PD-inspired	Uniform
4	PD-inspired	Sparse

Table 1: Simulation summary

Data	Method	Accuracy	
		avg.	s.d.
PD 1.0	MV	84.32	-
	MPA	91.43	-
Synthetic Uniform	MV	85.09	0.52
	MPA	90.12	0.52
Synthetic Sparse	MV	76.55	0.46
	MPA	85.92	0.60
PD-inspired Uniform	MV	89.26	0.47
	MPA	97.38	0.28
PD-inspired Sparse	MV	82.72	0.56
	MPA	94.36	0.33

Table 2: Mention pair accuracy results. Each simulated scenario is randomly generated 10 times (summary is in terms of average result and standard deviation)

The simulations are then generated by extracting from each gold label a number of 'crowdsourced labels' produced by (simulated) annotators with varying degrees of ability. We considered a range of simulated scenarios, all sharing the following settings:

- 10 distinct annotators per mention and 20 distinct mentions per annotator. The annotators receive random mentions to annotate.[6]
- Each annotator is assigned randomly a **profile**. The profiles indicate the sensitivity of the annotators with respect to discourse old and new. For example, the (DO 0.8, DN 0.9) profile indicates that, given a mention whose true class is DO, the annotator has 0.8 probability of getting it right; and of 0.9 for DN. We considered both profiles reflecting the actual profiles of players in *Phrase Detectives* (Chamberlain, 2016) and synthetic profiles.
- 5 choices for the annotators to choose from for each mention: the correct label, the DN

[6] This equal load reflects work distribution as found in microtask crowdsourcing rather than in games such as *Phrase Detectives*, where a few players do most of the work (Chamberlain, 2016).

PD 1.0	Method	MUC			BCUB			CEAFE			Avg. F1
		P	R	F1	P	R	F1	P	R	F1	
Singletons included	MV	95.18	69.44	80.30	95.53	78.79	86.36	79.04	95.12	86.34	84.33
	MPA	92.87	86.07	89.34	94.79	88.56	91.57	90.53	94.27	92.36	91.09
	Stanford	65.55	59.70	62.49	79.83	74.54	77.09	77.74	85.41	81.40	73.66
Singletons excluded	MV	95.18	69.44	80.30	93.36	46.05	61.68	64.23	55.17	59.35	67.11
	MPA	92.87	86.07	89.34	88.46	72.83	79.89	79.65	76.32	77.95	82.39
	Stanford	65.55	59.70	62.49	51.09	39.16	44.33	41.44	49.02	44.91	50.58

Table 3: The quality of the coreference chains on the PD 1.0 dataset

PD 1.0	Method	P	R	F1
Non Referring scores	MV	82.98	20.00	32.23
	MPA	75.14	66.67	70.65

Table 4: Non-referring scores for the PD 1.0 dataset

label (without including it twice if this is the correct label), and the 3 (or 4 if the correct label is DN) incorrect DO antecedents situated closest to the mention.

The range of options considered in the simulation is specified by two aspects: the sensitivity from the annotator profiles and the distribution of the errors they make. We use the following two profile types:

- Synthetic profiles: 5 profiles covering a wide range of abilities (DO 0.8, DN 0.9), (DO 0.7, DN 0.8), (DO 0.4, DN 0.5), (DO 0.3, DN 0.4), (DO 0.2, DN 0.3). The profiles roughly correspond to two experts and three novices whose class sensitivities are relatively close – with extra mass associated with DN because this class is generally easier compared to DO.

- *Phrase Detectives* inspired profiles: from the PD annotators who annotated more than 10 DO and 10 DN mentions (thresholds set to have a minimum confidence) we extracted a total of 89 profiles. This gave us much more interesting sensitivity pairs compared to the ones from the synthetic profiles, i.e., contrasting class abilities – see Figure 2.

We also considered a range of ways in which annotators may make mistakes:

- Distribute the errors uniformly random given the remaining mass (1 - sensitivity)

- Distribute the errors in a sparse manner, i.e., assume that some errors will be more likely than others. This can be achieved by drawing randomly from a 4-dimensional (4 = number of errors) uniform Dirichlet for each mention. The annotator probabilities over the 5 choices will then consist of their sensitivity, and the error distribution normalized with respect to the remaining mass.

The settings just discussed lead to 4 simulations summarized in Table 1.

3.2 Evaluation 1: Mention Pair Accuracy

We use MPA to link each mention with the most likely label based on the posterior of the true label indicators. We then assess the accuracy of the inferred mention pairs against the gold standard, i.e., the agreement with the gold mention pairs. In this task the proposed model is compared against a majority vote baseline where each mention is paired with the most voted label.[7]

The evaluation is conducted on the crowdsourced annotated PD 1.0 dataset and on simulated data generated from the CONLL-2012 test set. The results, summarized in Table 2, indicate the mention pairs inferred by our model (MPA) obtain a much better level of agreement with the gold mention pairs, compared with the output of the majority vote (MV) baseline. MV implicitly assumes equal expertise among the annotators, which has repeatedly been shown to be false in annotation practice (Poesio and Artstein, 2005; Passonneau and Carpenter, 2014; Plank et al., 2014).

3.3 Evaluation 2: Silver Chain Quality

After the mention pairs have been inferred using MPA, producing the coreference chains – we will henceforth refer to the coreference chains thus ob-

[7]Throughout the paper we report the best majority vote result after 10 random rounds of splitting ties.

CoNLL 2012 Test Dataset			MUC			BCUB			CEAFE			Avg. F1
Simulation	Method		P	R	F1	P	R	F1	P	R	F1	
None	Stanford		89.78	73.88	81.06	83.93	59.22	69.44	73.87	60.57	66.56	72.35
Synthetic Uniform	MV	avg.	88.27	86.00	87.12	73.92	70.81	72.33	70.62	76.73	73.55	77.67
		s.d.	0.38	0.35	0.36	0.83	0.52	0.62	0.49	0.60	0.50	0.47
	MPA	avg.	90.92	91.97	91.44	75.51	80.14	77.75	81.98	78.81	80.37	83.19
		s.d.	0.48	0.36	0.41	1.20	0.67	0.93	0.74	1.16	0.93	0.75
Synthetic Sparse	MV	avg.	81.99	79.01	80.47	65.91	62.64	64.23	60.61	68.00	64.09	69.59
		s.d.	0.32	0.43	0.38	0.45	0.51	0.40	0.39	0.24	0.28	0.32
	MPA	avg.	87.90	88.24	88.07	70.67	73.91	72.25	74.62	73.63	74.12	78.15
		s.d.	0.47	0.44	0.45	0.96	0.65	0.77	0.75	0.93	0.82	0.66
PD-inspired Uniform	MV	avg.	91.84	88.28	90.02	80.94	74.19	77.41	75.13	84.93	79.73	82.39
		s.d.	0.36	0.49	0.42	0.61	0.84	0.66	0.88	0.55	0.72	0.58
	MPA	avg.	97.42	97.20	97.31	91.61	91.53	91.57	93.87	94.58	94.23	94.37
		s.d.	0.27	0.28	0.27	1.05	1.28	1.15	0.67	0.61	0.63	0.68
PD-inspired Sparse	MV	avg.	86.86	81.70	84.20	74.26	65.42	69.56	65.45	78.51	71.39	75.05
		s.d.	0.49	0.54	0.51	0.63	0.48	0.52	0.67	0.55	0.60	0.53
	MPA	avg.	94.86	94.09	94.47	85.24	84.42	84.83	87.52	89.91	88.70	89.33
		s.d.	0.32	0.36	0.34	0.70	0.75	0.71	0.60	0.49	0.54	0.52

Table 5: The quality of the coreference chains on the CoNLL-2012 test set. Each simulated scenario is randomly generated 10 times (summary reported in terms of average result and standard deviation)

tained as **silver** coreference chains[8] – is a straightforward clustering task: we simply follow the link structure from the pairs. In this Section we assess the quality of the silver chains using standard coreference metrics – in particular, the Extended Scorer introduced in (Poesio et al., 2018) which extends the official CONLL scorer to include in the evaluation system-predicted singletons and non referring expressions, both of which are annotated in *Phrase Detectives*; when singletons and non-referring expressions are not considered, the Extended Scorer is identical to the official scorer.

As in the previous experiment, the evaluation is conducted on the crowdsourced annotated PD 1.0 dataset and on simulated data generated from the CONLL-2012 test set. We compare silver chains produced using our MPA model, using MV, and using the Stanford deterministic coreference system (Stanford) (Lee et al., 2011). To run the latter on PD 1.0, we used the default annotators of the CoreNLP toolkit (Manning et al., 2014) to supply the information required by the coreference system and switched off the post-processing to output singleton clusters; for the CONLL-2012 data we set the $dcoref.replicate.conll = true$ to run exactly the same method as Lee et al. (2011). On both datasets we evaluated on gold mentions.

Table 3 summarizes the results on the crowdsourced annotated PD 1.0 dataset. The silver chains obtained using our MPA model are of a far better quality than those of baseline alternatives such as MV and Stanford. Note also that even the simple MV baseline built from crowdsourced annotations yields much better chains compared to a standard coreference system such as the Stanford system. This underlines the advantage of crowdsourced annotations for coreference over automatically produced annotations. In Table 4 we present the scores of MPA and MV on cases of non referring. In this case, as well, the probabilistic model substantially outperforms the MV baseline.

In Table 5 we present the results obtained on simulated data from the CONLL-2012 test set. The results follow a similar trend to those observed using actual annotations: a much better quality of the chains produced using the mention pairs inferred by our MPA model, across all the simulated

[8]Our use of the term 'silver standard' should not be confused with the other common use of standard generated out of automatic annotations.

			MUC			BCUB			CEAFE			Avg. F1
Simulation	Method		P	R	F1	P	R	F1	P	R	F1	
None	Gold		78.40	73.40	75.80	68.60	61.80	65.00	62.70	59.00	60.80	67.20
	Stanford		79.87	63.67	70.86	71.63	47.85	57.37	58.55	48.08	52.80	60.34
Synthetic Uniform	MV	avg.	78.67	67.51	72.65	67.68	51.41	58.41	59.59	52.63	55.89	62.32
		s.d.	0.87	0.73	0.13	1.48	0.99	0.27	0.62	0.64	0.34	0.22
	MPA	avg.	78.27	70.21	74.02	66.67	56.06	60.90	61.44	54.92	57.99	64.30
		s.d.	0.64	0.57	0.23	1.03	0.81	0.31	0.50	0.51	0.27	0.24
Synthetic Sparse	MV	avg.	77.95	64.45	70.55	66.80	47.21	55.29	57.64	49.64	53.34	59.73
		s.d.	0.75	1.18	0.52	1.17	1.75	0.89	0.64	1.10	0.76	0.72
	MPA	avg.	77.99	68.82	73.11	66.01	53.76	59.25	60.25	53.46	56.65	63.01
		s.d.	0.55	0.51	0.25	0.93	0.84	0.42	0.43	0.38	0.27	0.28
PD-inspired Uniform	MV	avg.	78.99	68.44	73.33	68.42	52.86	59.63	59.99	54.04	56.85	63.27
		s.d.	0.58	0.60	0.14	0.92	0.75	0.38	0.79	0.27	0.44	0.27
	MPA	avg.	78.35	72.39	75.25	67.65	59.89	63.53	62.32	57.70	59.92	66.23
		s.d.	0.24	0.34	0.15	0.58	0.42	0.20	0.21	0.38	0.25	0.14
PD-inspired Sparse	MV	avg.	78.72	65.46	71.47	67.99	48.43	56.55	58.29	51.33	54.59	60.87
		s.d.	0.70	0.68	0.33	1.43	0.96	0.55	0.57	0.62	0.53	0.45
	MPA	avg.	78.34	71.47	74.75	67.43	58.11	62.42	61.90	56.77	59.22	65.46
		s.d.	0.44	0.62	0.20	0.83	0.94	0.26	0.27	0.49	0.28	0.23

Table 6: Results of a state of the art coreference system trained on silver chains obtained in different ways. Each simulated scenario is randomly generated 10 times (summary is in terms of average result and standard deviation)

scenarios. Furthermore, the MV baseline achieves better chains compared to the Stanford system in 3 out of 4 simulation settings, again showcasing the potential of crowdsourced annotations.

3.4 Training on Silver Chains

In this Section we assessed the viability of using the (silver) chains extracted from crowdsourcing as an alternative to gold chains when training a state of the art coreference system. Concretely, we train the best-performing current system Lee et al. (2017) on chains produced using our MPA model, the MV baseline and the Stanford deterministic system (Lee et al., 2011) (used mainly for calibration, i.e., an alternative baseline that's not based on crowdsourced annotations). We also include the results obtained using actual gold chains.

The results are in Table 6. Across all simulated scenarios, the silver chains produced by our MPA model obtain the closest performance to training on gold chains, and the best result is only 1 percentage point less than the result with gold chains. Again, the MV chains lead to better performance than those obtained using a system (Stanford). These results, once again, indicate the utility of crowdsourced annotations for coreference tasks.

3.5 Traditional Crowdsourcing Tasks

In this Section we show that MPA is state of the art also on traditional crowdsourcing datasets, where annotations fall into general classes that are consistent across the annotated items. This evaluation was done on the datasets (*WSD*, *RTE* and *TEMP*) introduced by Snow et al. (2008) and widely used as benchmarks in the literature on annotation models (Hovy et al., 2013; Carpenter, 2008).

We compare the results against a majority vote baseline and two well-known state of the art models: a Bayesian version of the Dawid and Skene (1979) (DS) model and MACE (Hovy et al., 2013). We implement DS ourselves using variational inference, while for MACE, we simply report the published results. As in Hovy et al. (2013) the assessment is done in terms of accuracy against the gold standard. The results, presented in Table 7, indicate the proposed model achieves performance on par with the state of the art.

4 Related Work

To our knowledge, this is the first paper proposing a model of crowdsourced annotations for coreference. We did draw inspiration however from existing mention pair models of coreference and traditional models of annotation.

The so-called mention pair model is one of the early machine learning approaches to coreference resolution, made popular by Soon et al. (2001). The model is based on a two step procedure: a classification step which identifies the coreferent mention pairs, followed by a clustering step which builds the coreference chains from the aforementioned pairs. The diversity of mention pair models comes from the distinct approaches taken for each of the two steps (Hoste, 2016). Although we follow a similar two step procedure, there are also important differences. Our way of identifying the mention pairs is completely unsupervised, and relies entirely on the crowdsourced annotations. Furthermore, we pair every mention with only one label, reducing the second step of clustering mention pairs into appropriate coreference chains to a simple grouping task guided by a unique path which arises from the pairs.

All existing probabilistic models of annotation (Dawid and Skene, 1979; Smyth et al., 1994; Raykar et al., 2010; Hovy et al., 2013; Passonneau and Carpenter, 2014) assume the annotations fall into a general set of classes that is consistent across the annotated items. This is clearly not the case in a coreference resolution task, a limitation we had to address. We first transformed the annotations into a series of (per label) binary decisions, approach often referred to, in the multi-class classification literature, as the binary relevance method (Tsoumakas and Katakis, 2007; Madjarov et al., 2012). The transformation avoids modeling the sparse labels directly. We further exploited the fact that the annotations fall into a general set of classes and assumed the inter-label decisions are the result of the class-dependent ability of the annotators.

5 Conclusions

Crowdsourced annotations are an increasingly popular alternative to expert annotation. Even so, their viability for coreference annotation had not been explored so far. This paper is a first step to filling this gap. We introduced a mention pair-based approach to aggregating crowdsourced anaphoric annotations and assessed the quality of the inferred pairs, of the post-hoc constructed coreference chains, and the viability of using the inferred chains as an alternative to gold chains when training a state of the art coreference system. Throughout the experiments, the model introduced was superior to baseline alternatives such as majority vote and chains obtained automatically using a coreference system, across both genuinely crowdsourced and simulated coreference datasets. Furthermore, even the annotation-based baseline achieved results consistently better than those obtained by automatic coreference resolvers, strengthening the case for using crowdsourced annotations to create coreference datasets.

	RTE	TEMP	WSD
MV	90.00	93.00	99.00
MACE	93.00	94.00	99.00
DS	93.00	94.00	99.00
MPA	93.00	94.00	99.00

Table 7: Accuracy on standard crowdsourced data

Acknowledgments

Paun, Chamberlain, Juntao and Poesio are supported by the DALI project, funded by ERC.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596. An early version of this paper has been circulating since 2005 as "Kappa[3] = Alpha (or Beta)". This version is still available from the ARRAU website.

Breck Baldwin and Thomas S. Morton. 1998. Dynamic coreference-based summarization. In *Proceedings of the Third Conference on Empirical Methods for Natural Language Processing*, pages 1–6.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 47–57.

David M. Blei and Michael I. Jordan. 2006. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J. Venhuizen, and Johannes Bjerva. 2017. The Groningen Meaning Bank. In *The Handbook of Linguistic Annotation*, chapter 18, pages 463–496. Springer.

Bob Carpenter. 2008. Multilevel Bayesian models of categorical data annotation. Available at http://lingpipe-blog.com/lingpipe-white-papers.

Jon Chamberlain. 2016. *Harnessing Collective Intelligence on Social Networks*. Ph.D. thesis, University of Essex, School of Computer Science and Electronic Engineering.

Jon Chamberlain, Massimo Poesio, and Udo Kruschwitz. 2016. Phrase detectives corpus 1.0: Crowdsourced anaphoric coreference. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2016)*, Portoroz, Slovenia.

Nancy A. Chinchor. 1998. Overview of MUC-7/MET-2. In *Proceedings of the* 7^{th} *Message Understanding Conference (MUC7)*.

Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Alexander Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28.

Paul Felt, Robbie Haertel, Eric K. Ringger, and Kevin D. Seppi. 2014. MOMRESP: A Bayesian model for multi-annotator document labeling. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik.

Ralph Grishman and Beth Sundheim. 1995. Design of the MUC-6 evaluation. In *Proceedings of the 6th Conference on Message Understanding*, MUC6 '95, pages 1–11, Stroudsburg, PA, USA. Association for Computational Linguistics.

Anupam Guha, Mohit Iyyer, Danny Bouman, and Jordan Boyd-Graber. 2015. Removing the training wheels: A coreference dataset that entertains humans and challenges computers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347.

Veronique Hoste. 2016. The mention-pair model. In *Anaphora Resolution: Algorithms, Resources and Applications*. Springer.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. Event coreference for information extraction. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, CONLL Shared Task '11, pages 28–34, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and SašO Deroski. 2012. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084–3104.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association of Computational Linguistics*, 3(1):405–418.

Thomas S. Morton. 1999. Using coreference for question answering. In *Proceedings of the Workshop on Coreference and Its Applications*, CorefApp '99, pages 85–89, Stroudsburg, PA, USA. Association for Computational Linguistics.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. 2004. An annotation scheme for information status in dialogue. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2004)*.

Rebecca J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon.

Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Computational Linguistics*, 2:311–326.

Barbara Plank, Dirk Hovy, and Anders Sogaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

Massimo Poesio, Jon Chamberlain, and Udo Kruschwitz. 2017. Crowdsourcing. In *Handbook of Linguistic Annotation*. Springer.

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase Detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Intelligent Interactive Systems*, 3(1).

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the NAACL Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC)*, pages 11–22, New Orleans.

Massimo Poesio, Sameer Pradhan, Marta Recasens, Kepa Rodriguez, and Yannick Versley. 2016a. Annotated corpora and annotation tools. In *Anaphora Resolution: Algorithms, Resources and Applications*. Springer.

Massimo Poesio, Roland Stuckardt, and Yannick Versley. 2016b. *Anaphora Resolution: Algorithms, Resources and Applications*. Springer, Berlin.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL 2012)*, Jeju, Korea.

Sameer S. Pradhan, Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. Ontonotes: A unified relational semantic representation. In *International Conference on Semantic Computing (ICSC 2007)*, pages 517–526.

Nguyen Quoc Viet Hung, Nguyen Thanh Tam, Lam Ngoc Tran, and Karl Aberer. 2013. An evaluation of aggregation techniques in crowdsourcing. In *Web Information Systems Engineering – WISE 2013*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. 2010. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322.

Arndt Riester, David Lorenz, and Nina Seeman. 2010. A recursive annotation scheme for referential information status. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)*, pages 717–722.

Aashish Sheshadri and Matthew Lease. 2013. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of the 1st AAAI Conference on Human Computation (HCOMP)*, pages 156–164.

Padhraic Smyth, Usama Fayyad, Michael Burl, Pietro Perona, and Pierre Baldi. 1994. Inferring ground truth from subjective labelling of Venus images. In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'94, pages 1085–1092, Cambridge, MA, USA. MIT Press.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Josef Steinberger, Massimo Poesio, Mijail Kabadjov, and Karel Jezek. 2007. Two uses of anaphora resolution in summarization. *Information Processing and Management*, 43(6):1663–1680. Special issue on Summarization.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 2007:1–13.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing. *Handbook of Natural Language Processing and Machine Translation. Springer*.

Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. 2009.

Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22*, pages 2035–2043. Curran Associates, Inc.

Zhiping Zheng. 2002. Answerbus question answering system. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 399–404, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

A Supplementary Parameter Estimation Details

This Section gives supplementary details for the derivations involved in the parameter estimation process.

In Equation (10) we derive the complete conditional of the class-specific true label likelihood:

$$\begin{aligned}
p(\pi_h|...) &\propto p(\pi_h|a,b)\prod_{i,m}^{I,M_i} p(c_{i,m}|\pi_h)^{I(z_{i,m}=h)} \\
&\propto \pi_h^{a-1}(1-\pi_h)^{b-1}\times \\
&\times \prod_{i,m}^{I,M_i} \pi_h^{I(c_{i,m}=1)I(z_{i,m}=h)} \\
&\times (1-\pi_h)^{I(c_{i,m}=0)I(z_{i,m}=h)} \\
&\propto \pi_h^{a-1+\sum_{i,m}^{I,M_i} I(z_{i,m}=h)I(c_{i,m}=1)} \\
&(1-\pi_h)^{b-1+\sum_{i,m}^{I,M_i} I(z_{i,m}=h)I(c_{i,m}=0)} \\
&\propto \mathsf{Beta}(a+\sum_{i,m}^{I,M_i} I(z_{i,m}=h)I(c_{i,m}=1), \\
&b+\sum_{i,m}^{I,M_i} I(z_{i,m}=h)I(c_{i,m}=0))
\end{aligned} \tag{10}$$

The corresponding variational distribution has the same form, i,e, $q(\pi_h|\lambda_h,\eta_h) = \mathsf{Beta}(\lambda_h,\eta_h)$. Taking the expectation of the natural parameters of the above distribution gives the variational update formulas for the class-level true label likelihood expressed in the paper. For example, for the λ_h variational parameter we have:

$$\begin{aligned}
\lambda_h &= a+\sum_{i,m}^{I,M_i} I(z_{i,m}=h)E_q[I(c_{i,m}=1)] \\
&= a+\sum_{i,m}^{I,M_i} I(z_{i,m}=h)\phi_{i,m}
\end{aligned} \tag{11}$$

Similar steps were taken to derive the variational parameters associated with the sensitivity α and specificity β.

In Equation (12) we derive the complete conditional associated with the positive outcome of the true label indicator:

$$\begin{aligned}
p(c_{i,m}=1|...) &\propto p(c_{i,m}=1|\pi_{z_{i,m}})\times \\
&\times \prod_{n}^{N_i} p(y_{i,m,n}|\alpha_{jj[i,m,n],z_{i,m}}) \\
&\propto \pi_{z_{i,m}} \prod_{n}^{N_i} \alpha_{jj[i,m,n],z_{i,m}}^{I(y_{i,m,n}=1)} \\
&(1-\alpha_{jj[i,m,n],z_{i,m}})^{I(y_{i,m,n}=0)} \\
&\propto \exp\{\log\pi_{z_{i,m}}+ \\
&\sum_{n=1}^{N_i} I(y_{i,m,n}=1)\log\alpha_{jj[i,m,n],z_{i,m}}+ \\
&I(y_{i,m,n}=0)\log(1-\alpha_{jj[i,m,n],z_{i,m}})\}
\end{aligned} \tag{12}$$

The corresponding variational distribution has the same form, i,e, $q(c_{i,m}|\phi_{i,m}) = \mathsf{Bern}(\phi_{i,m})$. Taking the necessary expectations leads to the update formula expressed in the paper. Concretely, we have:

$$\begin{aligned}
&\log\phi_{i,m} \propto E_q[\log\pi_{z_{i,m}}]+ \\
&+\sum_{n=1}^{N_i} I(y_{i,m,n}=1)E_q[\log\alpha_{jj[i,m,n],z_{i,m}}]+ \\
&+I(y_{i,m,n}=0)E_q[\log(1-\alpha_{jj[i,m,n],z_{i,m}})] \\
&\propto \Psi(\lambda_{z_{i,m}})-\Psi(\lambda_{z_{i,m}}+\eta_{z_{i,m}})+ \\
&+\sum_{n=1}^{N_i} I(y_{i,m,n}=1)[\Psi(\gamma_{jj[i,m,n],z_{i,m}})- \\
&\Psi(\gamma_{jj[i,m,n],z_{i,m}}+\mu_{jj[i,m,n],z_{i,m}})]+ \\
&+I(y_{i,m,n}=0)[\Psi(\mu_{jj[i,m,n],z_{i,m}})- \\
&\Psi(\gamma_{jj[i,m,n],z_{i,m}}+\mu_{jj[i,m,n],z_{i,m}})]
\end{aligned} \tag{13}$$

The update formula for the negative outcome of the true label indicator $\zeta_{i,m}$ is derived in a similar manner. Following the above derivations should also make it straightforward to expand the ELBO.

For completeness, we make a note of the digamma function $\Psi()$ – this is the first derivative of the $\log\Gamma$ function and can be computed using a Taylor approximation.